Early Operational Experience with the Cray X1 at the Oak Ridge National Laboratory Center for Computational Sciences

Buddy Bland, Richard Alexander
Steven Carter, Kenneth Matney, Sr.

Cray User's Group
Columbus, Ohio
May 15, 2003

Center for Computational Sciences

DOE's Advanced Computing Research Testbed

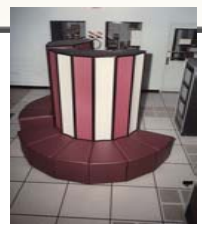
CCS Center for Computational Sciences

OAK RIDGE NATIONAL LABORATORY
U. S. DEPARTMENT OF ENERGY

Goals of the Center for Computational Sciences

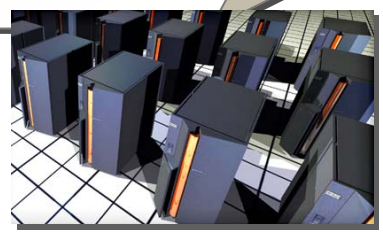
- Evaluate new computer hardware for science
- Procure the largest scale systems (beyond the vendors design point) and develop software to manage and make them useful
- Deliver leadership-class computing for DOE science
  - By 2005: 50x performance on major scientific simulations
  - By 2008: ***1000x*** performance
- Education and training of next generation computational scientists

ORACLE (1954–60)



Cray X-MP (1985)

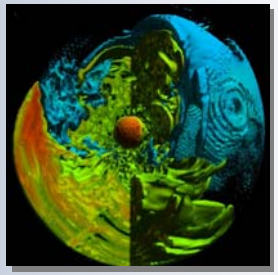
Intel XP/S 150 (1995)



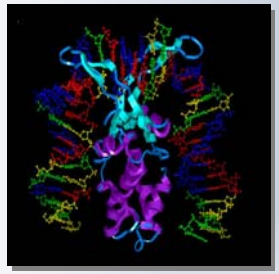
IBM Power4

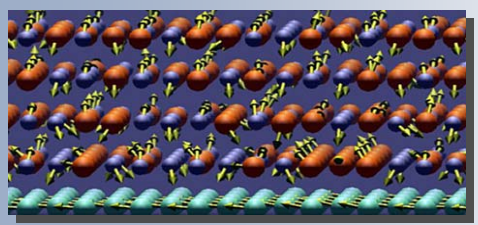
Focused on grand challenge scientific applications

Astrophysics

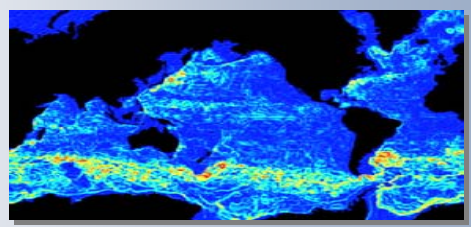


Genomes to Life



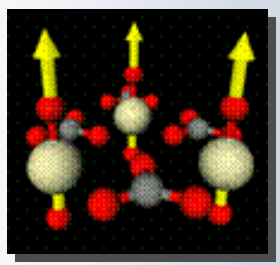


Nanophase Materials

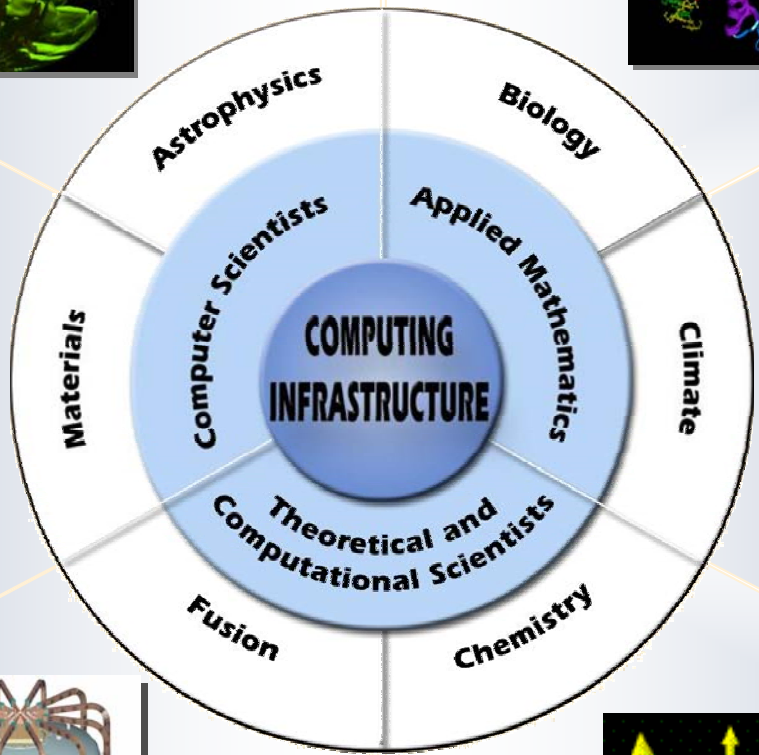


Climate

Fusion



Chemistry

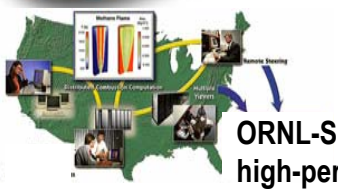


A decade of firsts (1992-2002)

1992
First Paragon XP/35
KSR1-64
CCS formed

1993
PVM used to create first International Grid

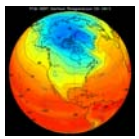
1994
PVM wins R&D 100



1995
Install Paragon XP/150
Worlds fastest computer
Connected by fastest network OC-12 to Sandia

First OSC TeraFLOP peak system

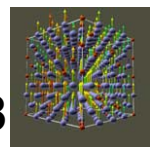
1996
ORNL-SNL create first high-performance computational Grid

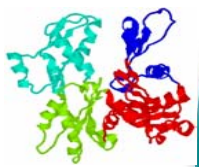


2000
Longstanding climate simulation milestone first met on CCS Compaq

1997
R&D 100 Award for successful development and deployment of HPSS

1998
Developed first application to sustain 1 TF



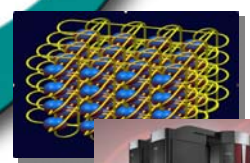


1999
NetSolve wins R&D 100
ATLAS wins R&D100

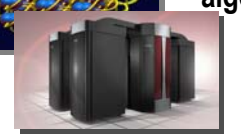
2001

First IBM Power 4
SciDAC leadership
Human Genome

2002



IBM Blue Gene CRADA to develop super scalar algorithms begins



Partnership with Cray on X1 begins

2003
Design changes for X2 based on ORNL-Cray partnership

Construction starts on new CCS building
World class DOE facility

Requirements drove the construction of a new world class facility capable of housing petascale computers

- Space and power for world class facilities
  - 40,000 ft^2 Computer Center
  - 36” raised floor; 18 ft. deck-to-deck
  - 8 megawatts of power (expandable)
- Office space for 400 staff members
- Classroom and training areas for users
- High ceiling area for visualization lab (Cave, Power Wall, Access Grid, etc.)
- Separate lab areas for computer science and network research
- Strong university partnerships

New State Funded Joint Institute for Computational Sciences for Academic Outreach

- **Special facilities for user access to ORNL terascale computing capabilities**
- **High-speed networking, forefront visualization tools**
- **Programs for grad students and postdocs**
- **Joint appointments**

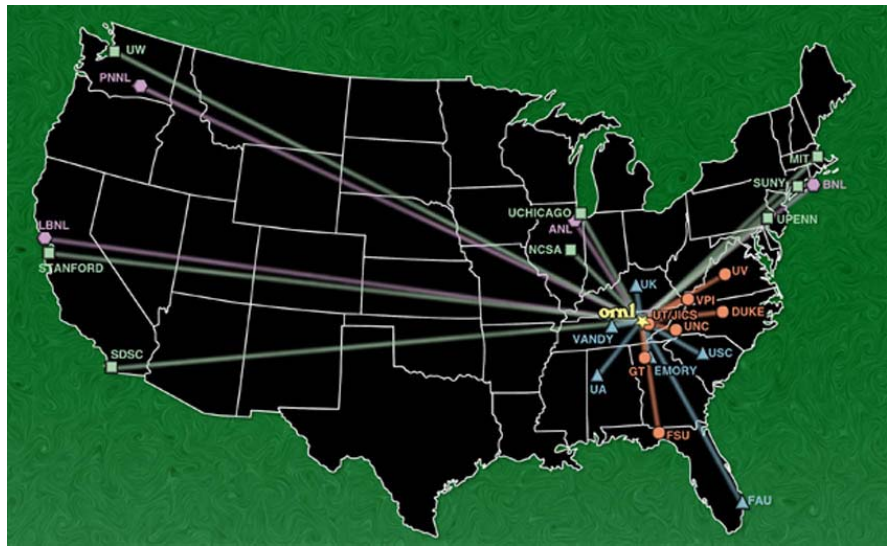
CCS Center for Computational Sciences

Computational Sciences Initiative:

- **Collaborative program between UT and ORNL to expand computational sciences**
- **UT invests ~1.3M for CCS fellowships and faculty release time**
- **UT invests in OC192, high speed network connection**

State-of-the-Art network connectivity to the Nation's principal Scientific Networks (ESnet, Internet 2)

School	Old Bandwidth to ORNL/DOE	New Bandwidth to ORNL/DOE
● UT-Battelle & Associated Universities		
University of Tennessee	155 Mbs	10,000 Mbs
JICS	1,000 Mbs	10,000 Mbs
University of Tennessee, Chattanooga	1.5 Mbs	1,000 Mbs
Georgia Tech	45 Mbs	10,000 Mbs
UNC	45 Mbs	622 Mbs
Duke	45 Mbs	622 Mbs
University of Virginia	155 Mbs	155 Mbs
Virginia Tech	155 Mbs	155 Mbs
Florida State University	155 Mbs	155 Mbs
▲ Other Southeastern Schools of Interest		
University of Alabama	45 Mbs	155 Mbs
Emory	45 Mbs	1,000 Mbs
University of Kentucky	45 Mbs	155 Mbs
University of South Carolina	45 Mbs	622 Mbs
Florida Atlantic University	45 Mbs	155 Mbs
Vanderbilt	45 Mbs	622 Mbs
■ Other Research-One Universities		
MIT	45 Mbs	622 Mbs
SUNY Stony Brook	155 Mbs	155 Mbs
University of Penn	622 Mbs	622 Mbs
University of Ill - NCSA	622 Mbs	10,000 Mbs
University of Calif - SDSC	622 Mbs	10,000 Mbs
University of Chicago	622 Mbs	622 Mbs
Stanford	622 Mbs	10,000 Mbs
University of Washington	100 Mbs	10,000 Mbs
● DOE National Labs		
LBNL	622 Mbs	622 Mbs
PNNL	155 Mbs	155 Mbs
BNL	622 Mbs	622 Mbs
ANL	622 Mbs	10,000 Mbs
LLNL	622 Mbs	622 Mbs
LANL	622 Mbs	622 Mbs



- ESnet – OC12
- Internet2 – OC192

CCS Center for Computational Sciences

OAK RIDGE NATIONAL LABORATORY
U. S. DEPARTMENT OF ENERGY

The CCS has a long history of evaluating new computer architectures.

- Over 30 different evaluations of beta or serial number 1 systems over the last 15 years

Current Systems at the CCS

CHEETAH – IBM Power 4

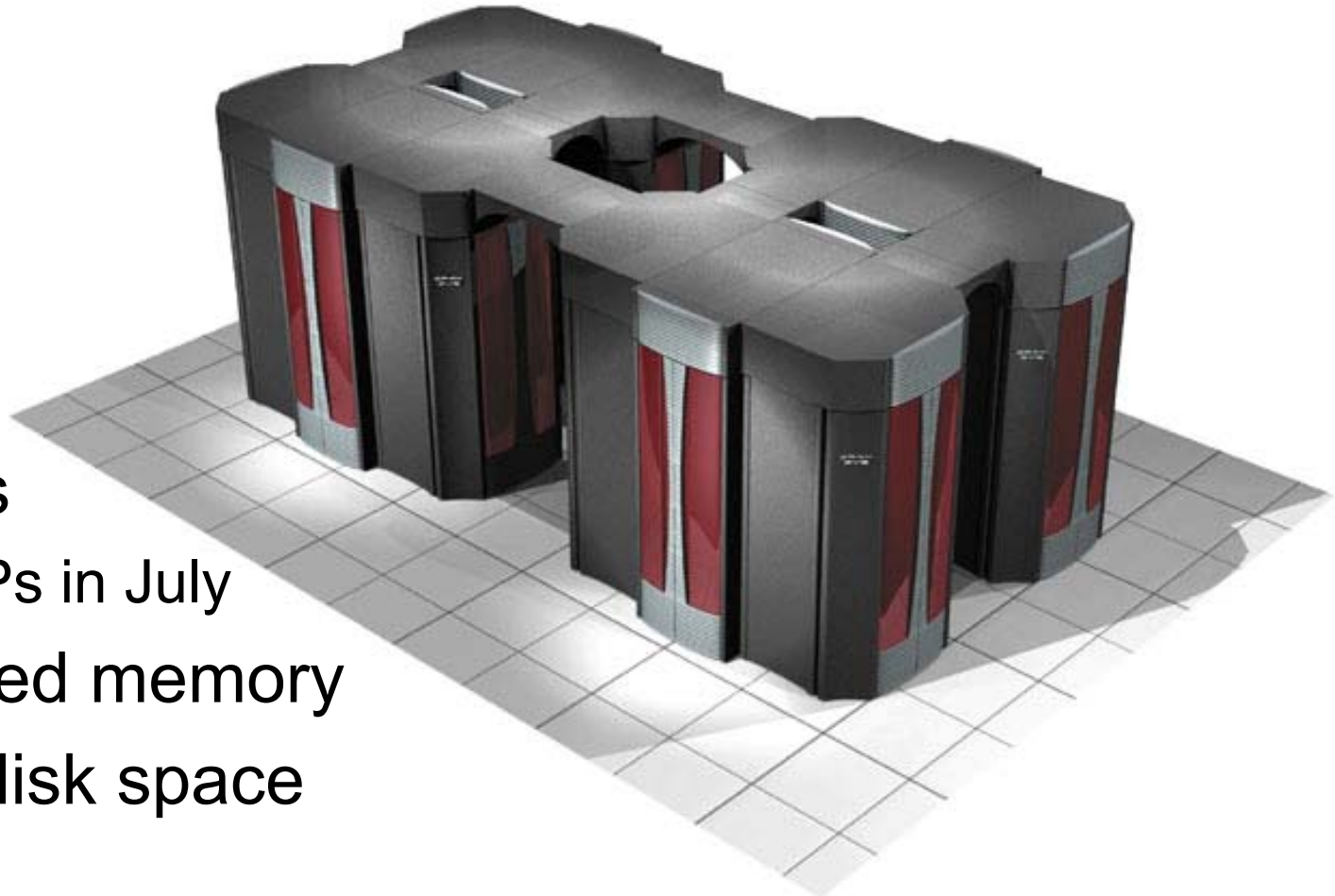
- 4.5 TFlops, IBM Power4 Regatta System
- 8th on Top500 list (June 2002)
- 27 computational nodes each with 32 processors
- 864 computational processors
- 5.2 GFlops/sec peak processor speed
- 1.2 TB total memory
- 40 TB of disk space in GPFS
- 400 TB archival disk storage HPSS
- Upgrade to Federation interconnect in 2003

Cray X1 - Phase 1
March 18, 2003

- 32 MSPs
  - 8 nodes, each with 4 processors
- 128 GB shared memory (4GB per MSP)
- 8 TB of disk space

400 GigaFLOP/s

Phase 2 – Summer 2003



- 256 MSPs
  - 128 MSPs in July
- 1 TB shared memory
- 20 TB of disk space

3.2 TeraFLOP/s

Experiences with the Cray X1

System Installation

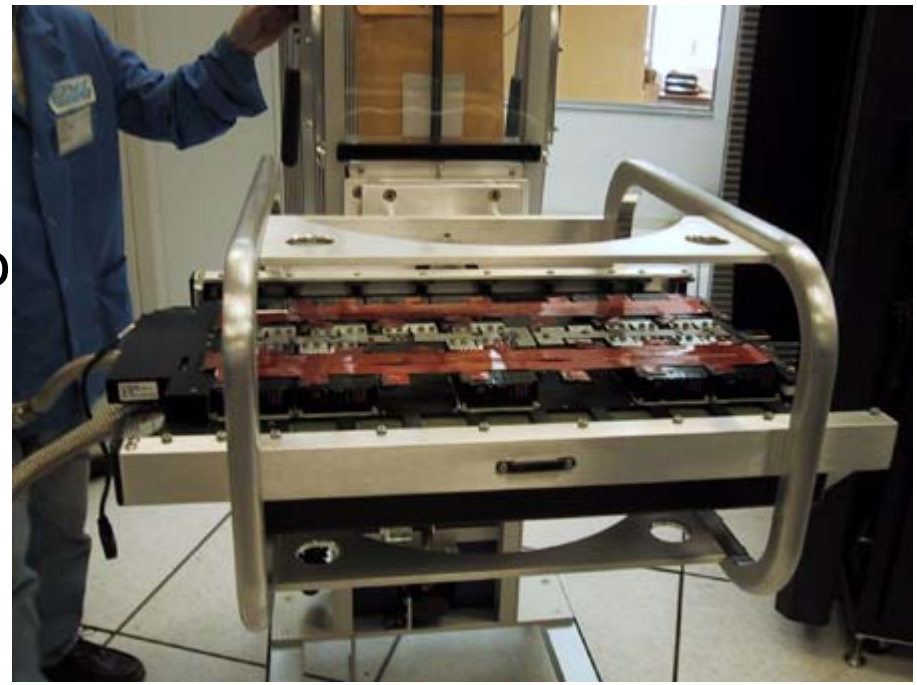
CCS Center for Computational Sciences

Installation was flawless!

- Cray site prep team had done their homework
- Very tight turn from a hallway into the computer room. Cray designed a custom dolly to be able to pivot the machine in place to make this turn
- 8” raised floor so the machine had to be jacked up to make the chilled water connections before it was set in place
- Powered on Wednesday, passed acceptance test on Friday
- We wish that 480 volt power was an option!

Hardware Experiences

- Two problems at hardware installation
  - Disk drive failure
  - One node board that had to be reseated
- Problems since
  - Failed power supply on node board
  - Hung L0 diagnostic controller – reboot fixed the problem



UNICOS/mp

- New operating system, so stability expectations were low
- Measured MTTI from Cray system is 86 hours
- However, this does not include “hangs” that necessitated a reboot instead of a crash
- Stability is actually better than we expected. We are in a software development stage and are able to get the work done
- Cray has been responsive in fixing problems, but there are still problems and we have had some performance regressions.

UNICOS/mp - psched

- Two problems
  - **`psched`** will hang and sometimes not dispatch jobs. Requires a reboot to clear the problem.
  - Essentially not integrated with PBSpro at this time. PBSpro job start is not an atomic operation. Can lead to race conditions. Integration is coming later this summer.

UNICOS/mp –Check this slide

- We are having mystery crashes
- Probably one user, but not yet isolated to a single code
- Some evidence that points to one user, but nothing yet definitive.
- Ongoing problem for the last 10 days.
- Need maturity in dump analysis to track these types of problems.

Integration into CCS environment

- No DCE available so we used the Army version of Kerberos 5 supplied by Cray. We have not been able to compile the PD version yet.
- Minor configuration problems
  - **`sendmail`** would not delivery mail externally
  - Configuring the fiber channel disk logical volumes for good performance turned out to be a challenge

Important problems – TCP performance

- We have had some network problems using TCP for NFS.
- Very slow performance. Mounts from X1 were timing out.
- We have reconfigured it to use UDP, but that bypasses the `TCP-ASSIST` feature of the CNS that does packet assembly.
- Filling a GigE pipe takes a significant fraction of an MSP.
- Small packet performance is poor and performance for all packet sizes needs to be improved. Small packets are a fact of life. The performance has to be improved.

Important problem – Site selectable authentication methods

- UNICOS/mp provides basic **`passwd`** file
- Need to allow sites to have more options using *something like* PAM from Linux
- This is complicated since UNICOS/mp does not support dynamic libraries. Most of the code is in **`libc`** and changes require relinking all programs
- Discussions underway with Cray on how to solve this problem. Feedback from other sites is needed to find a general solution.

Improvement needed in Programming Environment

- Compile time is very slow, but has improved from the first version.
- Cross compile environment will be available for Sun soon, Linux later?
- Slow NFS makes this problem worse.
- Compiler is stable and generates good code. The PE group has made good progress. Still work to be done.

Future Plans

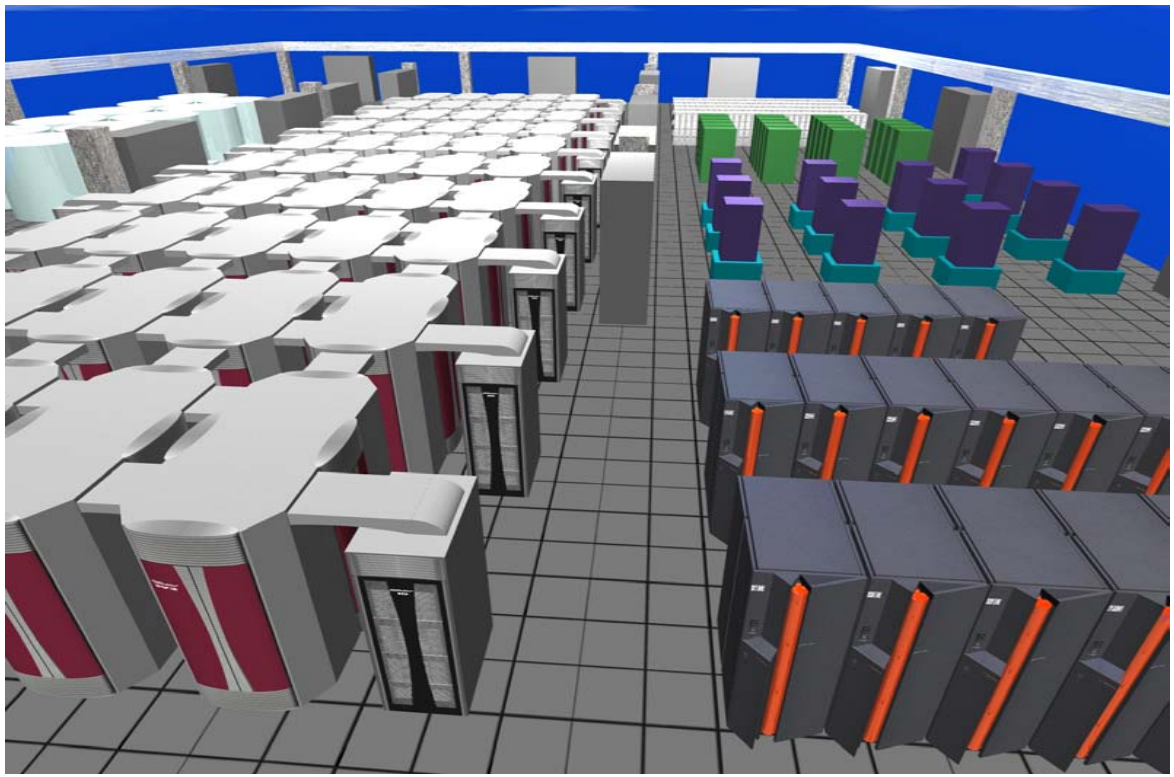
Where do we go from here?

Phase 3 – Fourth Quarter 2003

- 640 MSPs
- 2.5 TB shared memory
- 50 TB of disk space

8 TeraFLOP/s

Phase 4 – FY 2005



Proposed 40 TF configuration
in ORNL's new computer center

- Upgrade to 3200 MSPs
- 13 TB shared memory
- 100+ TB of disk space

40 TeraFLOP/s

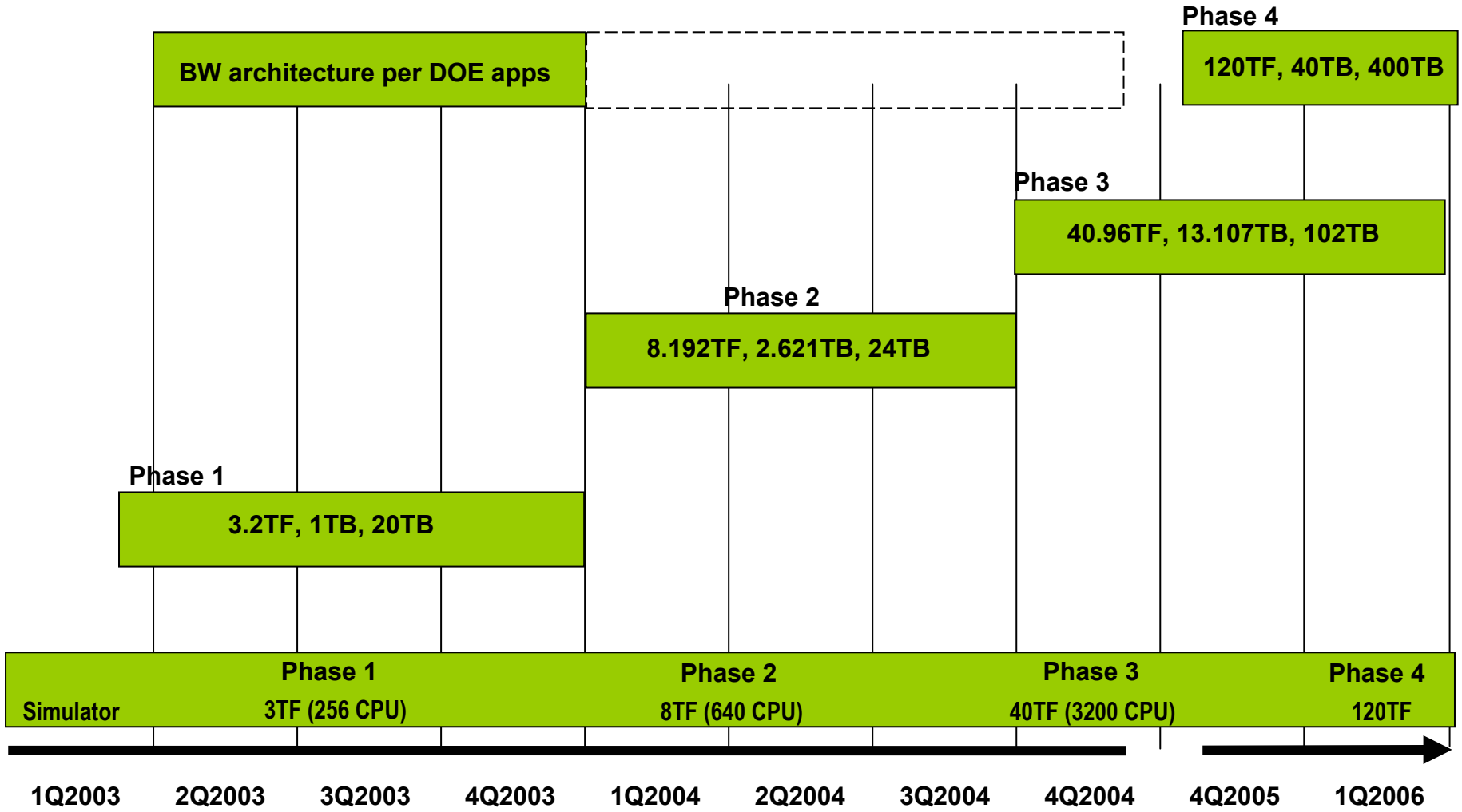
Red Storm

- Sandia – Cray collaboration
- CCS is collaborating with Sandia on applications
- CCS will likely purchase a test system in 2004
- Long term – Need a plan for software integration of Red Storm and Black Widow software efforts

Phase 5 – Cray "Black Widow"

- Next generation processor
- Improved performance
- Improved price/performance
- Current plan calls for a system in excess of 100 TeraFLOP/s in late FY2006
- Important opportunity to influence the design choices based on the needs of DOE apps

Cray X1/Black Widow
4 phase evaluation and deployment



Early Operational Experience with the Cray X1 at the Oak Ridge National Laboratory Center for Computational Sciences

Buddy Bland
BlandAS@ornl.gov

Richard Alexander
alexar@ccs.ornl.gov

Steven Carter
scarter@ornl.gov

Ken Matney
matneykdsr@ornl.gov